DOCUMENT RESUME

ED 227 181 UD 022 603

AUTHOR	Inbar, Michael
TITLE	Images or Aberrations? Human Judgment and Insight as Reflected In Current Regression Analyses.
SPONS AGENCY	National Council of Jewish Women, New York, N.Y. Research Inst. for Innovation in Education.
PUB DATE	Apr 82
NOTE	73p.
AVAILABLE FROM	Michael Inbar, Department of Sociology, Hebrew University, Mount Scopus, Jerusalem 91905, Israel (write for price).
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Evaluative Thinking; Higher Education; *Judgment Analysis Technique; Multiple Regression Analysis; Performance Factors; Policy Formation; Predictor Variables; *Statistical Bias; Student Characteristics; *Validity
IDENTIFIERS	*Beta Weights; *R2 Values

ABSTRACT

In contrast to linear models of human judgment developed for predictive purposes which are characteristically insensitive to the exact values of the weights utilized in them, the Linear Multiple Regression Models used for policy capturing are assumed to reflect significant aspects of the subjects' judgmental policies. This latter kind of modelling is therefore justified only to the extent that its underlying assumption is found to be true. Its validity depends, however, on both that of the models' beta weights, and of R2 as a twin measure of the subjects' cognitive control/consistency and of one's success in capturing the judges' policy. Although the problematic nature of beta weights is well known, the present study shows that R2 is no less a problematic measure. It is shown that the way in which self-insight is typically elicited may induce a demand-response effect; furthermore, traditional data analysis and comparison methods appear to be inconsistent. In the study, eight subjects were presented with 72 profiles of perspective undergraduate students and were asked to judge future academic performance, based on information regarding each student's sex, age, ethnic oigin, and socioeconomic and educational background. The results indicate that current policy capturing research and findings cannot be accepted at face value. (Author/GC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED227181

IMAGES OR ABERRATIONS?

Human Judgment and Insight as Reflected In Current Regression Analyses*

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Michael Inbar

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Michael Inbar

Department of Sociology

The Hebrew University of Jerusalem

1982

CG 022 603

* The present study has been sponsored by the National Council of Jewish Women (NCJW) Research Institute for Innovation in Education, under the auspices of the Barbara and Morton Mandel Chair in Cognitive Social-Psychology and Education, the Hebrew University of Jerusalem. The work reported is part of a study of human judgment generically labelled Project MIBS (Models of Implicit Belief Systems). Requests for reprints should be addressed to the author, Department of Sociology, Hebrew University, Mount Scopus, Jerusalem 91905, Israel.

ABSTRACT

In contrast to the linear models of human judgment developed for predictive purposes which are characteristically insensitive to the exact values of the weights utilized in them, the Linear Multiple Regression (LMR) models used for policy capturing are assumed to reflect, partly through them, significant aspects of the subjects' judgmental policies. This latter kind of modeling, be it for research, for providing cognitive feed-back, for training or for assessing the subjects' self-insight, is therefore justified only to the extent that this underlying assumption is found to be so. Its validity depends, however, on both that of the models' β's, and of R^2 as a twin measure of the subjects' cognitive control/consistency and of one's success in capturing the judges' policy. The problematic nature of the beta weights has long been known. The present study, based on 8 subjects, shows that R^2 is no less a problematic measure. Moreover, with the data of 4 of these subjects in one case, and with that of 7 in the other, it is shown that the way in which self-insight is typically elicited may induce a demand-response effect; additionally, the traditional manner of analyzing and comparing these data appears to be flawed by a grave inconsistency. Together, these results indicate that current policy-capturing research and findings cannot be accepted at face value. A list of threats to the validity of these models and their application is offered. The likelihood that studies which have disregarded their possible relevance and impact on the results obtained may be reporting misleading findings is stressed. In conclusion the dependence of the justification of the policy-capturing endeavor on confronting these problems is pointed out.

IMAGES OR ABERRATIONS?

Human Judgment and Insight as Reflected In Current Regression Analyses

INTRODUCTION

Background

Modeling human judgement by means of linear multiple regressions (LMR) has become a standard procedure. This technique is common to the tradition of research that Hammond and his colleagues (e.g. Hammond, McClelland and Mumpower, 1980) have labelled Social Judgment Theory (SJT), and to that which can be traced to Meehl's (1954) work through such studies as those of Dawes (1971), Goldberg (1970) and Hoffman (1960). These traditions of research have produced an extensive body of findings which has been surveyed in a number of articles and reviews, including Slovic and Lichtenstein (1971), Dawes and Corrigan (1974), Hammond, Stewart, Brehmer and Steinmann (1975) and Brehmer and Hammond (1977). A salient theme in this literature as a whole is a keen interest in three related issues: capturing the judges' policies; determining the degree of the judges' insight into these policies; and comparing the validity of the decisions made by the LMR models of the judges with the validity of the decisions made by the judges themselves. The overall pattern of the findings reported is rather consistent. It suggests three main conclusions which can be stated and concisely illustrated as follows: 1) LMRs yield efficient models of the judges' policies (e.g. "a simple linear model will normally permit the reproduction of 90-100% of [the clinical judges'] reliable judgmental variance", Goldberg, 1968, p. 491; see also Hoffman, 1968, pp. 59-60; Einhorn, Kleinmuntz and Kleinmuntz, 1979, p.468). 2) Judges lack insight into their judgmental policies (specifically, "... a number of studies, varying in the number of cues

that were available [have shown that] three cues usually sufficed to account for more than 80% of the predictable variance in the judges' response ... One type of error in self-insight has emerged in all of these studies. Judges strongly overestimate the importance they place on minor cues (i.e. their subjective weights greatly exceed the computed weights for these cues) and they underestimate their reliance on a few major variables.", Slovic and Lichtenstein, 1971, p. 684; see also Hobson, Mendel and Gibson, 1981, pp.181-182, for the same conclusion based on the average use of four rather than three cues). And 3) the models are usually more valid than the actual judgments from which they were originally derived (and, thus, "One is left with the conclusion that humans may be used to generate inference stategies but that once the strategy is obtained, the human should be removed from the system and replaced by his own strategy!", Dudycha and Naylor, 1966, pp. 127; quoted by Goldberg, 1970, p. 431).

These generalizations have been occasionally qualified. Thus, it has been shown that man can outperform his model (Libby, 1976a). Man is also able to simultaneously use at least eleven cues under laboratory conditions (Phelps and Shanteau, 1978). A configural model may at times provide a better fit than a linear one (Hoffman, 1968; Einhorn, 1970, 1971; Libby, 1976b). On the issue of insight, Cook and Stewart (1975) found that statistical and subjective weights were in good accord, in contradiction to the bulk of previous research (cf. Schmitt and Levine, 1977, p 16); Schmitt (1978) also found that statistical and subjective weights correlated highly, although the statistical weights were slightly but significantly superior to the subjective ones as predictors of the subjects' judgments; Gray (1979, p. 30) using a single cue experimental paradigm found that the judges' insight was limited. He concludes that his findings and the available evidence supports the composed generalization that "people's effectiveness in predicting uncertain events exceeds their ability to express insight into their prediction process".

There have also been developments in the opposite direction, notably the sharpening of the proposition that man could usefully be replaced by his model. The clearest expression of this trend is found in Dawes and Corrigan (1974) and Dawes (1979) who have provided a rationale for using "improper" linear models with equal weights. Analytical considerations as well as nearly a decade of empirical work suggest that this recommendation cannot be lightly dismissed (Einhorn and Hogarth, 1975; Dawes, 1979; Camerer, 1981).

On the whole, then, the three generalizations noted above appear to be well substantiated and widely held (Slovic, Fischhoff and Lichtenstein, 1977; Hammond, McClelland and Mumpower, 1980, Hogarth, 1980; Shapira,1981); they are moderately qualified, but more in the spirit of setting their practical limits than of challenging their veracity.

The Focus of the Present Study

The three issues of capturing the judges' policy, determining their insight, and comparing the validities of man and his models, are often empirically interrelated. Analytically, however, they are distinguishable. The present research focuses on the two first issues, and deals only incidentally with the third.

In well known papers, Hoffman (1960; 1968) and Darlington (1968) have warned against the pitfalls of identifying paramorphic models or their parameters with the psychological processes being modeled. Schmitt and Levine (1977) have convincingly reiterated this warning. The heart of the methodological argument is that currently there is no single statistical index for reliably and, therefore, meaningfully measuring (capturing) policies. This argument has been so compellingly presented, both theoretically and empirically, that the

question arises as to the reason(s) behind the continued use of LinR models for this purpose, or for the related one of providing subjects with a yardstick for assessing their self-insight. The reason is certainly not theoretical in some substantive sense, for the two schools of thought which originated this line of reasearch explicitly disavow such a goal. Hoffman (1968) already went on record some fifteen years ago to express his distrust of the results that paramorphic modeling was documenting. Dawes' work has brought these reservations and the connection between substantive modeling and the issue of statistical robustness to their logical end. He views linear models as variously weighted, additive indices, justified by the degreee of their statistical efficiency, but not presumed to paramorphically or otherwise model whatever aspect or level of the judgmental process itself (Dawes and Corrigan, 1974; Dawes, 1979). Similarly, Hammond and his colleagues (Hammond, McClelland and Mumpower, 1980, pp. 61, 71, 105, 136) take great pain to make it clear that SJT has nothing to say about human judgment per se. They repeatedly emphasize that the logic of the cognograph and cognitive feed-back approach rests on what this school of thought has been willing to extrapolate from what is in essence a learning rather than a judgmental paradigm of research. This emphasis is in line with the fact that the methodological pitfalls noted above concern primarily the estimation of the β's and that Brehmer and his colleagues (Brehmer and Qvarnstrom, 1976; Brehmer, Hagafors and Johansson, 1980) have shown that capturing the judges' policy involves precisely the estimation of these quantities rather than that of the less controversial correlation coefficients. The foregoing position is also consistent with the prevalent use of the lens model equation as reformulated by Tucker (1964) in terms of correlation coefficients only.

Under these conditions of explicit lack of theoretical grounds or patronage, the explanation for the continued "modeling" or capturing of policies and

the inferences about the judges' insight that they currently sustain can only be guessed. The best available clue is perhaps provided by the rationale offered by the authors of a recent study. In the process of reviewing some of the major weaknesses inherent in LMR modeling, Einhorn, Kleinmuntz and Kleinmuntz (1979, pp.467-468) remark that "... indeterminacy in estimating weights when cues are correlated (Darlington, 1968), parallels the organism's difficulty in this matter...[and] The inconsistency and random error in judgment, resulting from the lack of cognitive control in executing one's strategy ... is explicitly defined and measured within regression procedures." They then go on suggesting that when LMRs are viewed in the light of such characteristics, "... they seem neither arbitrary nor ad hoc nor devoid of psychological content. Furthermore, the great success of such models in a wide variety of tasks strongly suggests that some fundamental characteristic of judgment has been captured... . The success alluded to is presumably the remarkable ability of LMRs to explain most of the explainable variance of judges' responses, an ability which, in contrast to the problematic beta weights, has indeed remained largely unquestioned. In this perspective, inferences about the judges' self-insight may be based on the discrepency between the small number of cues with which a LMR typically reproduces a judge's decision and the more numerous ones that the judges report having taken into consideration, rather than on the problematic comparison of β weights. Under the circumstances, the conclusion will remain unchanged: man lacks self-insight.

The aim of the present study is to document that to the extent that the foregoing line of reasoning serves in this form or some related one as an implicit or explicit justification for the continued attempts to capture man's judgmental policies according to established practices, it is problematic in its

own right as well. Specifically, the number of profiles typically used in LMR studies of human judgment and the fact that policies are not static (Brehmer, 1978; Bucuvalas, 1978), on the one hand, and the manner in which self insights about the cues and weights used by the subjects are elicited, on the other, combine to undermine and often invalidate such a rationale. The heart of the problem is that the pivotal measures of consistency or cognitive control (R) and the subjective information collected about subjects' reliance on cues can be shown to be at times artifactual. Pessimism about man's cognitive consistency and/or insight into his policies may nonetheless be warranted. The two findings just alluded to suggest, however, that it is unsafe to infer this from current LMR analyses --and under most circumstances, neither is it wise to expect these analyses to be able to help remedy whatever cognitive shortcomings man demonstrably has.

A peculiarity of the research to which we now turn to document the two results just noted should be pointed out. The findings were accidentally documented in a study which addressed different issues. Because by their logic these findings are independent of many specific characteristics of a typical policy-capturing study, they do not require a specially designed study for their demonstration. For the sake of convenience, they are presented with the data of the study in which they were originally documented.

THE STUDY

Overview

The research under consideration involved eight subjects who acted as individual judges. This investigation followed a pre-test which was conducted with the aim of applying the standard LMR paradigm of analysis to the judgments of both individuals and groups, the purpose was to investigate mismatches between

certain findings and intuition (e.g. the differential number of cues in the subjects' models and in their retrospective reports) by means of process-tracing. As a result of this background, the study had a complex design with which we need not concern ourselves here. Suffice it to note that each subject performed a number of judgmental tasks over a period of about three weeks. The findings and analyses which will be discussed pertain to the first of these judgmental tasks.

Subjects

Six undergraduates, one accountant and one MD acted as subjects. They were recruited through personal connections, and selected after having been made aware that the experiment would last several weeks and might at times seem repetitive. The subjects expressed their willingness to fully cooperate and were paid over twice the usual hourly rate (a lump sum); it seems likely that their motivation included an element of curiosity and of willingness to help provide data for a scientific study.

The Task

The judges were presented with a set of 72 profiles. Each profile, alledgedly of a prospective undergraduate student, was to be judged in terms of the likelihood (0-100) that the quality of the undergraduate work of this candidate would be compatible with future graduate work. There were 16 cues per profile providing the following information about each applicant: sex, age, ethnic origin, I.Q., high school graduation grade, socio-economic background, marital status, health, achievement expectations, nature of relations with high school teachers, time spent doing homework during last year of high school, fear of failure, living expenses arrangements, political activities, social connections with university staff members, and sociability. Some of the cues were given quantitative values (e.g., age), others were described by quasi-interval

labels (e.g., no, some or intimate social relations with university staff members). The cues were moderately interrelated, the average of the absolute value of their intercorrelations being .132; the correlations ranged from -.71 to +.62, with the bulk of them (112 out of 120) ranging from -.31 to +.24.

Procedure

The task was individually explained during a practice session with 3-4 profiles. It was then handed out to each subject to be performed at home; the completed assignment was typically returned within 48 hours. Upon completion of the experiment as a whole each subject was individually debriefed.

Results

Table 1 presents the beta weights and the multiple correlations pertaining to the equations of the eight subjects. The alternating rows, A, B, and C give the results of three possible modeling decision-rules: inclusion of the variables with a beta weight significant at the .05 level or better only (A); inclusion of all the variables which contribute at least 1% of explained variance to the equation (B); inclusion of all the variables which contribute any measurable amount of explained variance to the equation (C). A shared constraint is that the overall R^2 of each equation be significant at the .05 level or better.

INSERT TABLE 1 ABOUT HERE

Choice of Equation

Each of these decision rules (selected for purposes which will become clear as the analysis proceeds) can be criticized.[1] The point here is not to repeat arguments already made (Darlington, 1968) but to introduce the main discussion by illustrating with the present data the kind of differences which can result from making one choice rather than another. Thus, in the case at hand, the average number of cues utilized which obtains for the three decision rules A, B, and C, is 5.5, 7 and 14.7 respectively. A subject can therefore arbitrarily end up being categorized as utilizing relatively few or many cues, depending on the equation chosen; moreover, the beta weights which presumably capture his policy correspondingly change. These results highlight the fact that in any attempt to capture the policy of a judge one difficulty revolves around the lack of objective criterion for selecting the equation which presumably best describes this policy. Note that since it is now recognized that the stepwise multiple regression procedure recommanded by Darlington (1968) will often yield results which compound the problems attached to the interpretation of the beta weights (Gordon, 1968; Cohen and Cohen, 1975), this procedure is not by itself an acceptable solution. By the same token, but more generally, any procedure relying on the amount of explained variance for selecting without additional rationale or safeguard the most appropriate equation, and by implication the most descriptive beta weights, is questionable. The reason is that this criterion will usually lead to the indiscriminate choice of the equation with the greatest number of variables, whether these are relevant or not. This follows from the fact that in a multiple regression the addition of a variable, even if redundant or irrelevant, can never reduce the amount of variance explained; if the variable is utterly redundant it will have no effect; if it is utterly irrelevant, it may have no effect, but often will add a quantum of explained variance, however minute, to the equation owing to chance relationships; one of

the situations noted in footnote 1 can then arise. When the sample size is held constant, the purpose of the adjusted R^2 (which, as far as it is concerned, may decrease; see discussion below) is precisely to correct for the inclusion of unnecessary variables in this sense. Table 1 illustrates the effectiveness of this correction. Although the increase in R^2 between equations A and C is on the average +5%, the corresponding difference between adjusted R^2's is negative (-.002); this trend is even accentuated when equations B and C are compared. The misguided (and unparsimonious) strategy of including all the variables which contribute any measurable amount of explained variance in an equation has therefore been properly identified by the values of the adjusted R^2's.

Along such a line of reasoning, it could be argued that the most appropriate equation for describing the policy of a subject should be selected on the basis of the largest adjusted R^2. This is in fact the logic of the strategy advocated by Wonnacott and Wonnacott (1979). This suggestion which has an undeniable appeal, has the substantive disadvantage that all stepwise procedures share (Gordon, 1968) and that the application under consideration does not avoid (but could minimize, a point to which we shall return).

The issue of immediate interest, however, is that when such a strategy is followed, it underscores an often overlooked characteristic of multiple correlation coefficients. Specifically, the use of adjusted R^2's makes salient the fact that the difficulties involved in capturing the policy of a judge are even more severe than is commonly realized.

The Notion of Cognitive Control

To put the foregoing in a concrete context, consider the observation that the policies of judges change during task performance (Brehmer, 1978; Bucuvalas,1978).

A reasonable question is to ask whether the change not attributable to unreliability takes the form of ad hoc applications of procedures as the need arises (Brehmer and Kuylenstierna, 1980) or of a more systematic change of policy over time, perhaps as a result of the processes of chunking and habituation.

A simple way to begin the investigation of this question is to split a sample of judgments according to their sequential order. If there should be a systematic over time change in policy, and if the split is adequately made, the equations developed within each new subsample should exhibit an improved fit over that found in the overall equation, within limits of sampling fluctuations. Operationally, therefore, one would expect the individual, and in any case, the average of the R^2's of the equations developed within the properly split subsequences of judgments to be greater than the R^2's of the equations developed on the whole sequence. Conversely, if no systematic change in policy takes place over time, no such expectation should be entertained.

The simplest possible sequential split is to separate the judgments into two equal groups, in our case two subsamples of 36 profiles, according to the order in which they were processed. If there should be in the present task only one major change in policy, and if it should typically take place about half way during task execution (as process tracing data suggests this might be roughly the case), this procedure, admittedly a gross approximation, should nonetheless help cast some light on the nature of policy change and policy routinization over time.

Table 2 presents the equations of the subjects developed in such a manner. In the interest of space, only one class of equations is presented. The equations are those which correspond to the decision rule which yields the highest adjusted R^2's in Table 1, decision rule B. The findings in Table 2 and the discussion which follows apply equally, however, in the case of the omitted

equations.

Insert Table 2 About Here

In terms of their β's, the twin equations are clearly very different from one another; similarly, they differ very much from the comparable equation developed on the whole sample for the same subject. Note, in particular, the not uncommon shift of variables, as well as change in signs, which occurs between the two sequential subsets of judgments.

In the light of the unstability of the beta weights noted earlier, including, as we have just seen, in the case of relatively slight variations of definition of the same equation (see Table 1), these results are neither surprising, nor necessarily indicative of any substantive process.

The values of the R^2's clearly suggest, however, that the policies during the first and the second half of the judgmental task were distinctly different. In terms of summary measures, the average of the R^2's in the two object subsamples is .78, as compared to .66 in the case of the parallel coefficient for the single equations of Table 1; moreover, in every single case the former average is greater than the R^2 of the corresponding equation developed on the whole object sample (see Table 2).

These values could be misleading; this is not unlikely owing to the combined effect of sample size and number of predictors in the new equations. The adjusted R^2's which correct for these parameters (see Table 2, column 20) suggest, however, that this is not the case. Although the adjusted values are noticeably reduced, the finding remains unchanged; the averages of the adjusted values of R^2's corresponding to those in the previous paragraph are, indeed, .70

and .62 respectively.

This finding has clearly potential implications for work on the modeling of policies and for the determination of the judges' insight into them. Because of the importance of these implications, it is prudent to double check the results. One way to accomplish this is to compare the results just obtained with those produced by splitting up the samples randomly. Table 3 presents the summary results of this analysis on subsamples divided by the odd even method; the relevant data from Tables 1 and 2 are included for comparative purposes.

Insert Table 3 About Here

Two results are of interest. The first is that by this standard as well, the evidence is that the subjects systematically used different policies during the first and second half of the task. The averages R^2's in the sequential subsamples is .78 versus .73 in the case of the randomly split subsamples; the adjusted R^2, .70 versus .66, respectively, corroborate these results (see the two penultimate rows at the bottom of Table 3).[2]

The second is that the average across subjects of the adjusted R^2's for the randomly split subsamples (.66) is greater than the corresponding average for the single equations(.62). This result which holds systematically true within subjects as well (with one exception, subject number 8, see Table 3, columns I-2 and III-2), makes salient the often disregarded fact that the adjusted R^2's may fail to adequately correct for variations in number of predictors and sample size. It is instructive to take a closer look at the reason for this failure.

It's source can be traced to the nature of the formula that is assumed to correct for variations in the two foregoing parameters (ie. to correct for "shrinkage"). This formula which adjusts for degrees of freedom has several related forms (see, Cohen and Cohen, 1975, pp. 106-107. Nie et al, 1975, p 358; Green and Tull, 1970, p.351). Because of the transparence of its structure, consider the form found in Wonnacott and Wonnacott (1979, p.181):

$$\overset{*}{R^2} = \left[R^2 - \frac{k}{n-1} \right] \left[\frac{n-1}{n-k-1} \right] \qquad (1)$$

where $\overset{*}{R^2}$ = adjusted squared multiple correlation, R^2 = obtained squared multiple correlation, k = number of predictors in the equation, and n = sample size. For instance, equation I for subject 1 in Table 2 yields a R^2 of .82; with k=9 and n=36, the corresponding $\overset{*}{R^2}$ is accordingly,

$$\overset{*}{R^2} = \left[.82 - \frac{9}{36-1} \right] \left[\frac{36-1}{36-9-1} \right]$$

$$= .76 \qquad (2)$$

The logic of this adjustment becomes evident if we note that the expected R^2 (R^2_E) in a multiple regression where none of the predictors is actualy related to the dependent variable, i.e., where the true value of $R^2 = 0$, will nonetheless be equal on the average to

$$R^2_E = \frac{k}{n-1} \qquad (3)$$

This follows from the fact that one can get a perfect fit to n data points using n-1 different predictors, independently of any other consideration (cf. Green and Tull, 1970, p. 351). Thus, with k=9 and a sample of n=10, the expected

R^2 is 1.0, even though the actual relationship may be 0. The first parenthesis on the right hand side of equation (1) corrects for this overestimation of R^2 by substracting from it the quantity $k/(n-1)$. If the predictors do bear some substantive relationship to the dependent variable, this adjustment is overdone, however. The reason can be seen by assuming that the criterion is actually perfectly related to the predictors, that is, $R^2 = 1$. The first parenthesis on the right side of equation (1) then yields the value:

$$\left[1 - \frac{k}{n-1}\right] = \left[\frac{n-1}{n-1} - \frac{k}{n-1}\right]$$

$$= \left[\frac{n-1-k}{n-1}\right] \qquad (4)$$

which is necessarily smaller than one, while by hypothesis $R^2=1$. Under these conditions, we would nonetheless like equation (1) to yield the value $\overset{*}{R}{}^2 = 1$. To insure that this is the case, the quantity (4) must be appropriately adjusted. This can be achieved by multiplying it by its inverse --the operation that the second term on the right hand side of equation (1) performs.

The correction that equation (1) achieves is, however, approximate, for it is not possible to determine exactly the degree of overestimation of R (cf. Kerlinger and Pedhazur, 1973, p. 282). The approximate nature of the procedure is probably best illustrated by noting that $\overset{*}{R}{}^2$ can be negative in which case it is by convention reported as 0 (Cohen and Cohen, 1975, pp. 106-107). For instance, to use these authors' example, for $R^2 = .10$, $k=11$ and $n=100$, equation (1) gives $\overset{*}{R}{}^2 = -.0125$.

The crux of the matter, then, is that equation (1) is only an estimate of the likely value of R^2 in the population. It gives a useful indication of the probable effect of cross-validation on any given R^2 as a function of the degrees

of freedom available when it was estimated. However, as the data in Table 3 illustrate, this correction is insufficient in the case of a stepwise analysis, the reason is that this kind of analysis affects the degrees of freedom by surreptitiously increasing the number of k's involved in the evaluation procedure, a difficulty which has led to the suggestion of various heuristic safeguards (Kerlinger and Pedhazur, 1973, pp. 282-283; Cohen and Cohen, 1975, p. 107; Wonnacott and Wonnacott, 1979, pp. 186-187). There is no evidence, however, of their being used in research on human judgment, despite the fact that the cautious use of stepwise multiple regressions does present advantages, and the observation that, whether critically applied or not, this procedure is commonly used for developing the models of the judges. Even more importantly, there appears to be a compartmentalization regarding the use of adjusted and unadjusted squared multiple correlations. While $\overset{*}{R}^2$ is increasingly reported in recent research, this is done as an indication of the likely effect of cross-validation on R^2, rather than for the purpose of better assessing cognitive control. Indeed, the central, and in many studies the only, measure of the concepts of cognitive control and consistency remains the uncorrected R^2.

This brings us to the heart of our present concern.

One important implication of the foregoing elaboration is that the nature of R^2 highlights the fact that in its unadjusted form the magnitude of R^2 is in part a direct function of the values of k and n. An often overlooked consequence of this relationship between R^2 (and, as just noted $\overset{*}{R}^2$ in stepwise analysis) and these parameters, is that the pitfalls attached to the direct interpretation of the explained variance as a measure of the strength of a relationship are not without resembling these found in the case of χ^2. Indeed, both types of measures reflect not only the strength of a relationship, but also the size of the sample involved in estimating it. In the case of χ^2, the

larger the sample, the greater the apparent relationship, while for multiple correlations, the larger the sample, the smaller it appears to be, other things being equal. (Incidentally, it is of interest to note that an informal survey shows that sophisticated researchers quite familiar with LMR techniques tend to have mistaken intuitions about the nature and direction of this effect of sample size on R^2). Another, more important difference is that the effect of sample size on the multiple correlation coefficient is for all practical purposes bounded. As the sample size increases, the expected shrinkage of this coefficient for any given number of predictors diminishes in direct relation to $k/(n-1)$, while in the case of χ^2 the sample size's effect remains undamped.

Let us now refocus our attention on Table 3.

The dependence of the magnitude of R^2 on the values of k and n and the fact that for small ratios of $k/(n-1)$ (i.e. for few predictors and large samples) the effect of these parameters may become negligible, suggest that the kind of findings reported in Table 3 ought to be interpreted in the light of the answers to two questions. The first is whether the artifactual effects illustrated in this Table are likely to be typical in LMR research on human judgment. The second concerns the practical implications of these artifacts.

Because equation (1) shows that $\tilde{R}^2$ is a function of k, n, and R^2, the answer to the first question depends on the magnitude of these quantities in empirical research. One estimate (Hammond, McClelland and Mumpower, 1980, pp. 132, 197) is that the typical values of k and n lie, respectively, between 5 and 8, and 20 and 50. With regard to R^2, Camerer (1981) found that the average $\sqrt{R^2}$ in 13 studies was .74; Shapira's (1981) survey of 22 (mostly different) studies yields the value.78. Slovic and Lichtenstein (1971) provide a more differen-

tiated estimate. They note that the $\sqrt{R^2}$'s they examined were in the .70's for complex, real-life judgments, while they were in the .80's and .90's for the more artificial, laboratory-type judgmental tasks. The following discussion integrates these estimates of the sizes of $\sqrt{R^2}$'s by preserving the distinction that Slovic and Lichtenstein made on the basis of their detailed review of the literature.

With this in mind, Table 3 justifies two conclusions. The first is that the trends documented on the basis of the subsamples of n=36 each, and the average number of cues in the equations developed on them of 6.6 (see column I, 3), are unlikely to be atypical. The second, is that owing to the size of the R^2's ($\sqrt{R^2} = \sqrt{.73} = .85$, on the average, see table 3, column I,1), the magnitude of the artifacts is probably more representative of that found in laboratory-type research, than in studies involving complex, real-life judgmental tasks or issues. Because the magnitude of the error is an inverse function of that of R^2, the size of the artifact will be greater in the latter case. The extent of the expected difference is illustrated in Table 4.

Insert Table 4 About here

For the sake of legibility, the relevant data have been organized into consecutive submatrices. The headings of these submatrices, $\overset{*}{R^2} = .30$ to $\overset{*}{R^2} = .80$, give the "true" values of explained variance that the selected values of the obtained R^2's listed in the corresponding submatrices yield by application of equation (1) or, equivalently, the values that the R^2's listed in the submatrices yield by a reverse application of equation (1) (i.e. $\overset{*}{R^2}$ given and

R^2 unknown) for combinations of values of k and n. For instance, the first entry in the first submatrix of Table 4 indicates that a R^2 of .48 obtained when k and n were 5 and 20, respectively, is likely to be in fact .30; conversely, a "true" squared multiple correlation of .30 is likely to have a value of .48 if it is estimated with 5 predictors on a sample of 20 profiles and, looking at the first entry of the last row of the same submatrix, a value of .37 if it is estimated with k = 5 and n=50; the discussion is notional and assumes that the values of $\overset{*}{R^2}$ in the headings of the submatrices are not biased by a stepwise procedure of estimation of the R^2's.

If for the purpose of clarity we trade precision for simplicity, and take $\sqrt{R^2} = .90$ as a point estimate to represent the typical range of values found in laboratory-type judgmental tasks and $\sqrt{R^2} = .75$ to represent that found in the more complex, real-life ones, the trends in Table 4 together with the main point made in the foregoing analysis, lead to the following conclusions.

Firstly, the effect of object sample size is greater across levels of k's than is that of the number of predictors across levels of n's. That is to say, in the range of values of k and n under consideration, a change in the size of the object sample tends to be more consequential than a change in the number of cues, whatever the level of cognitive control considered.

Secondly, for the laboratory-type tasks in which $R^2 = .81$ ($\sqrt{R^2} = .90$), and up, the magnitude of the artifact that relating undifferentially to k in the range 5 to 8 and to n in the range 20 to 50, may introduce in estimating a subject's cognitive control, or in comparing the findings produced by different studies, is on the whole relatively small. The last submatrix (with values of

R^2 ranging from .82 to .88, and $\overset{*}{R}{}^2$ =.80) shows, indeed, that when cognitive control reaches such a level, the maximum fluctuation in explained variance is 3% of explained variance when 8 rather than 5 cues (or vice versa) are used in a model, 5% when the object sample size changes from 20 to 50 (or vice versa) and 6% when both changes occur concurrently and additively (see the right hand diagonal of the submatrix under discussion). While these maximal values are not insignificant, the lesser magnitude of the other possible variations (some of which are 0 because of rounding necessities) may be regarded by the criterion suggested earlier as being of a magnitude where the advisability or not of distinguishing between discriminability and substantive significance is a matter of opinion.

Thirdly, for tasks in which subjects typically exhibit a $\sqrt{R^2}$ in the .70's, that is, to focus the discussion, where R^2 = .56 ($\sqrt{R^2}$ = .75), Table 4 shows that an empirical value of this magnitude is compatible with a true coefficient of cognitive control ranging from $\overset{*}{R}{}^2$ =.30 to $\overset{*}{R}{}^2$=.50. The submatrix headed by $\overset{*}{R}{}^2$ =.50 which best, and most conservatively, approximates the distribution of R^2's having the notional value of .56 of interest, indicates, moreover, that in this case the maximum fluctuation in explained variance is 8% (as compared to 3% in the previous case) when k varies from 5 to 8, 13% (versus 5%) when the object sample size varies from 20 to 50, and 16% (versus 6%) when both changes occur concurrently and additively.

It is probably noncontroversial to state that in this case neither the maximum potential magnitudes of the artifactual component of R^2, nor several of the lesser values it can have, can be safely disregarded; nor can the wide range of imprecision (.20 of "true" explained variance) regarding the magnitude

of the underlying coefficients of actual cognitive control (in this connection, another aspect of the effect of the size of R^2 may be noted; if we adhere strictly to the notional value of .81 discussed earlier, Table 4 shows that it is found in one submatrix only, that headed by the value $\overset{*}{R}{}^2 = .70$. That is to say, in the framework of the gross categories of Table 4, the imprecision shrinks in this case to 0, which by comparison with the previous value of .20 underscores the effect of the level of cognitive control on the pitfalls attached to the unguarded measurement of this concept).

Practically speaking, the seriousness of the foregoing artifacts depends on the magnitudes we have just documented; their consequentiality also depends, however, on the manner in which the artifacts tend to come about in actual research. That is to say, to assess their actual implications it is also necessary to have an idea of the conditions under which the quantities which determine the size of the artifactual component of R^2, namely k, n, and the size of R^2 itself, vary in empirical research in a potentially damaging fashion.

Consider first k. There seem to be at least two main ways in which the values of this parameter can undergo changes conducive to misleading inferences. The first is common in the situation where two equations developed for different judges on identical profiles (i.e. with the same set of supplied cues) and on an object sample of identical size, are directly compared. Under these circumstances, two subjects with identical true scores of cognitive control could, nay, are likely to end up being categorized as having different degrees of cognitive consistency, merely because their policies might involve a different number of variables, i.e. might require for their expression a different number of predictors in each equation. Similarly, the same subject studied on the same number of cases with profiles involving the same number of cues, but about a different

real-life substantive issue, could end up with different scores of cognitive consistency simply because of the number of cues he might happen to need to express -- exactly as well -- each of his policies. The second way in which k can change with confounding consequences is less insidious. The likely occurrence can be illustrated with a hypothetical study of transfer and generalization of the effect of cognitive feed-back that one may be tempted to carry out. In such a study, it could appear useful to design the criterion task with a different number of cues. If this second task should include more variables, and under the assumption of a monotonic relationship between number of cues provided and number of cues used in the judgmental task, a training session of this kind could be expected to produce a gain in cognitive consistency, if for nothing else, because of the direct relationship between the size of the artifactual component of R^2 and the number of predictors in a model.

Consider now sample size. We have seen above that the confounding effect of this parameter is greater than is the case for k. The pertinence, arbitrariness or accidental nature of the considerations which lead to the determination of the object sample size come also more readily to mind in this case owing to our sensitization to the issue of sample size in general. Thus, a probably shared experience is that these considerations include primarily the time available for the experiment, a typical value being one hour with students fulfilling a course requirement --with sometimes a follow up session of one more hour, often used for validating purposes and debriefing. When research money is available, the limiting consideration appears to be the anticipated information-processing capability of the paid subjects; sessions are then more likely to extend to 1½ or 2 hours, with as many additional sessions as necessary. For real life tasks the decisive factor is commonly the anticipated cooperation of the prospective judges -- which may in turn be a function of the social relations and/or the rapport of the researchers with them. That is to say, depending on the means

available to a researcher and to his perception of the patience of his subjects, the measured cognitive consistency of a judge can typically vary by as much as is made statistically possible by halving the size of an object sample, and even more than that if the combinations of sample size and of later validation by the split-half method are taken into account. On the whole, the modal range of sample sizes of n=20 to n=50 may express the manner in which the considerations and constraints just noted lead to the typical object sample sizes found in the literature.

Be this as it may, when the artifactual effect of sample size runs the risk of reaching the levels illustrated in Table 4, and when some of the controlling factors of this risk can be as irrelevant to the subjects' actual cognitive consistency as those we have just noted, it is clear that comparing levels of cognitive control across models without ascertaining the equality of the object sample sizes on which they were developed can be hazardous. As Table 4 shows, this hazard grows in direct proportion to the difference in sample sizes. It is noteworthy, however, that irrespective of the exact difference between n's, the probability of making misleading comparisons is always facilitated under the present circumstances by the fact that the sample size artifact operates in the same direction as does a seemingly compelling explanation. On statistical grounds, the greater the object sample size, the smaller R^2 is expected to be; similarly, from a substantive perspective, the greater the object sample size, the lower the degree of cognitive control one expects from the judges, and hence the smaller the intuitively expected R^2. A decrease in the size of R^2 observed in the context of a longer task is therefore likely to be interpreted as a decrease of the subject's cognitive control, despite the fact that part or all of such an effect is a necessary statistical outcome.

In short, not only are the artifactual values of R^2 we can expect to find in LMR research on human judgment likely to be at times of a magnitude we cannot safely disregard, but the manner by which these artifacts tend to come about suggests that they could be widespread.

Let us now summarize the main points of the foregoing discussion.

The pivotal observation is that the notion of cognitive control or consistency as measured by R^2 is ambiguous and problematic to an unexpected degree. For the typical range of values of k and n found in LMR research on human judgment, this measure is artifactually affected to a significant degree by variations in the values of these parameters and of the obtained R^2's. Table 3 (columns I,1 versus III, 1) concretely illustrates the operation of these confounding effects which range in this case from 0% to 21% of explained variance. The analyses based on Table 4 show that these values are recognizably close to theoretical expectations. Table 4 also permits one to phrase the problem differently. Thus, this Table indicates that stating that one subject exhibited a degree of cognitive control of, say, $R^2 = .58$ while that of another one was $R^2 = .71$, may simply convey information about the object sample sizes used in the two sessions. On the other hand, stating that two subjects have the same degree of cognitive consistency of, say $R^2 = .71$, may conceal the fact that despite the identity of this measure and the fact that the two judges were studied under identical conditions and on the same task (n=20, k=8), they actually have different true scores of cognitive control ($\overset{*}{R}^2 = .50$ and $\overset{*}{R}^2 = .60$, which translate into discrepencies of - 21% and - 11% respectively, with the measured score), a fact which is hidden by their different policies, involving in one case 8 cues and in the other 5.

These problems of interpretation of R^2 as a measure of cognitive control are compounded by the fact that Tables 2 and 3 suggest that subjects may be changing their policies over time in a systematic way. This points to the importance of determining the number of profiles which constitute a natural subset or block for capturing the policies between changes --an endeavor which could turn out to be idiosyncratic for at least some combinations of tasks by subjects. Because the range of typical sample sizes may involve one or more such natural blocks, the magnitude of R^2 is also likely to reflect the chance overlap, or lack thereof, between appropriately determined subsamples in this psychological sense, and the actual set of profiles on which a model happens to have been developed. The way the over time change combines with k,n, and the levels of obtained R^2 in affecting the artifactual component of this coefficient is a topic which deserves an analysis in its own right. Presently, it suffices to note that the larger the object sample, the greater the likelihood that R^2 will also be reduced on this account as a result of the mixture of policies in a single equation; note the implication that the dynamism of psychological processes and the statistical requirement of large n's may be working at cross purposes for the needs of modeling.

When we consider the effect of all the foregoing factors, either individually or in combination, the question evidently arises of the meaning and usefulness of R^2 as a measure of cognitive control, both to the researcher and to the subject. If we exclude a narrow range of values near the upper limit of its size, the absolute levels of this coefficient, as well as the possible changes observed in its values, are manifestly ambiguous to the point where its interpretation for either theoretical or applied purposes is at best uncompelling.

The analysis has assumed all along that there is some underlying human characteristic operationalized by R^2 which represents the subjects' true cognitive control and that R^2 presumably measures. It could be argued, of course, that there are no psychological grounds to expect the notion of cognitive control to be invariant across combinations of values of k and n, even in the modest range considered. This, however, is obviously not the manner in which R^2 has been used and reported in the literature on LMR modeling of human judgment. Such a view, moreover, raises with even greater acuity than do the indeterminacies and ambiguities discussed above the fundamental issue of what exactly is meant by the notion of cognitive control that R^2 assumedly measures--and that R^2 does not.

The crux of the matter, then, is that as a coefficient of cognitive control or consistency R^2 is a measure which in its current use for communicative purposes as well as in its practical applications, conveys information which is extremely difficult to interpret. Uncritically related to, this coefficient may therefore have little informative value; worse even, it runs the serious risk of being plainly misleading.

Although this is probably obvious, it may be useful to note that from a statistical viewpoint all that has been said above about R^2 as a measure of cognitive control or consistency, applies with equal strength to the interpretation of this coefficient as a measure of fit, that is, as a measure of success in capturing a judge's policy (cf. Hammond, McClelland and Mumpower, 1980, pp. 121, 149; Lane, Murphy and Marques, 1982). Indeed, whether the size of R^2 is attributed to the ability of the judge or to that of the modeler is irrelevant to the operation and magnitude of

the artifacts that we have discussed; clearly, this only affects the substantive process which is in danger of being misinterpreted.

Additional Results

We have seen earlier that according to one's choice of the equation which is deemed to represent a subject's policy, the number of cues in the model can widely vary (specifically from 5.5 to 14.7 cues on the average per equation in the empirical example discussed in Table 1). This problem, together with the ambiguities attached to the squared multiple correlation coefficient when it is used for the purpose discussed above, as well, as just noted, as in its use as a measure of success in capturing a judge's policy, lead to a self evident conclusion. It is that the grounds for determining whether a subject has or not self-insight into his policies are much less solid than is commonly assumed.

Nonetheless, once an equation with its R^2 is selected by whatever criterion, the assessment of the subjects' insight requires that data be obtained about what they feel their policies were. Several data gathering methods have been tried, yielding very similar results (Cook and Stewart, 1975). One of these methods, the scarcity scale, appears to have become standard procedure (Schmitt, 1978). The following discussion will focus on this scale.

Eliciting the Subjects Self-Insight

Consider briefly the structure of the scarcity scale method. The procedure consists of instructing the subjects to allocate 100 points among the cues in a manner which reflects these variables' relative importance in the set of judgments just completed. Response-wise, the subjects tend to carefully comply. They are scrupulous in two senses. Firstly, they are careful that the points allocated do add up to 100. Secondly, they attempt to allocate weights to all the variables that they have considered. Some subjects are scrupulous in the

first sense only, and do not allocate weights to "minor cues", although they are often hesitant about this and ask whether it is permissible. But many appear to interpret the instructions to include the second request as well, and attempt to allocate weights to all the variables, no matter how minute the discrimination they have to make and how uncertain they are about it. This part of the task is typically characterized by growing signs of hesitation, including erasures and the use as a last resort measure of some arbitrary rule for allocating a few points among the cues left over, or for redistributing the points so that every cue is included in the allocation.

By its logic, this extensive type of allocation clearly brings to mind the equations of type C in Table 1 where, it will be recalled, all the variables which can potentially enter into an equation are in fact forced into it by the nature of the decision rule. It was noted at the time that this is a misguided procedure. It seems peculiar therefore, that we should unwittingly put our subjects in a structural situation where they are forced, in fact, to do what should not be done, neither during model development, as we have noted earlier, nor in all probability during insight elicitation, as we have just indicated.

To put it differently, the findings obtained by means of the scarcity scale method could constitute a typical case of demand-response. This brings us to the second major topic of iterest in this paper.

To explore it, an alternative conception of the structure of the data which is required to assess self insight is desirable, for the situation just discussed is compounded by the fact that subjects keep insisting that their judgments are configural, and that they relate to clusters of variables rather than to individual cues. That is to say, we cannot simply rely on the use of

the most important of the subjects' introspective weights if we wish to go beyond the computational aspect of the demand reponse issue, and address the potentially even more consequential substantive one.

That subjects do often relate to combinations of cues is a common observation. For instance, in the case at hand, process tracing during the pretest showed that age, achievement expectations, high school grade and ethnic origin could be regarded by a subject as indicators of motivation, the concept he might say he was really trying to infer at that particular moment. More generally, the data also suggest that what the subjects actually attempt to do at this level falls into two categories of information processing: the interpretation of cues by means of other ones, and the inference of core variables or underlying "factors" (e.g. motivation) by means of subsets of cues, some of which act as stable indicators, while other vary as a result of the aforementioned interpretations.

In terms of our current concern, gathering data about this dual process, especially about the stable underlying clusters and assessing the weights of these "factors" in the judgments, is clearly one possible alternative way of the kind alluded to above to confront the subjects' claim and to evaluate the extent of their self-insight into their policies[3].

This approach to tapping the subjects' self-insight will yield a predictive index, similar to that produced by the standard elicitation of subjective data about the respondents' reliance on individual cues to which it is intended to be compared. This creates a problem in that there is no ready-made procedure for comparing the merits of indices. It would seem that a reasonable set of cri-

teria for the assessment of interest could include the following: the size of the correlation between the predictions based on the weights derived from the two types of self-insights and the actual judgments; the simplicity/complexity of the two indices; their theoretical construct validity; and their usefulness within the modeling process.

These criteria will be implicitly applied to the results of the analyses presented below.

Data Collection

The data were gathered by means of the instructions that standardly accompany scarcity scales. Specifically, after completion of the judgmental task the subjects were presented with a list of the 16 cues used in each profile and asked, in the case of insight about individual cues, to "Please allocate among the cues 100 points in a manner which reflects their relative weights in the judgments you made." After performing this assignment, the subjects were given a new list of the same cues with the following instructions: "Please consider again the 16 items of information. You may have related to clusters of them, rather than to individual items in making your judgments. If so, indicate next to each variable with which others you used it as a rule, by employing a common symbol for each grouping -- say, different numbers. If an item of information was used in several clusters, write down next to it the identifying symbol (number) of all the groupings to which it belonged".

After the completion of this task came the request to "Please give names to your groupings." This was followed by the concluding instruction to "Please allocate among these groupings 100 points in a manner which reflects the relative weight of each cluster in your judgments".

In all cases the foregoing instructions were sequentially handed out and explained to the subjects in a face to face session with an experimenter.

Subjects

Data are availabe for four subjects only. The idea of asking subjects to indicate how they had clustered the cues, if at all, and which weights they gave to these factors in their decisions emerged serendipitously during the debriefing of subject number 2 (who sparked the idea by spontaneously volunteering some of this information in the course of her justification of her objections to the questions asked about individual cues). The incipient idea quickly crystallized, and was operationalized in time for the debriefing of subject number 4 -- its first application. All subsequent subjects were asked the foregoing questions about clusters. Subject number 6, however, is an exception owing to a personal misfortune which interrupted her participation in the study just prior to being administered the self-insight scales. Hence the availability of data for subjects number 4,5,7, and 8 only.

Procedure

The general procedure for computing the predicted scores was the following. In the case of insight about individual cues, the z score of each cue (properly

signed -- the sign having been taken from the subjects' multiple regression, see Table 1[4]) was multiplied by its weight (zero, if the subject had disregarded the cue), and the predicted score for a given profiles was the sum of these products. In the case of clusters, the z scores of the cues constituting a grouping were first summed (after having been properly signed, as above), and each grouping was multiplied by its self-insight weight. The predicted score of a given profile was the sum of these products. Note that by this procedure all the cues defining a concept were given equal weight --obviously a gross (but conservative) oversimplification.

Findings

A. Prima facie validity of the "factors" indices. Columns I,II and III of Table 5 show that for three of the four subjects the judgments predicted on the basis of self-insight about the concepts inferred yield higher correlations with the actual judgments than is the case when the comparable predictions are made with data about individual cues. The tentative conclusion which emerges, therefore, is that information about clusterings of cues may be an alternative way of investigating the subjects' self-insight into their judgments, a way which in addition to being theoretically grounded appears to be empirically justified. Note that this conclusion is in essence similar to that reported by Cook and Stewart (1975), who also found that probing the subjects self-insight about interaction effects --albeit about individual cue utilisation rather than for configural concept inference (operationalized here by a simple additive index)--yielded predictions which tended to correlate higher with the actual judgments than did alternatively derived predictions.

Insert Table 5 about here

A related question is that of self-insight with regard to the substantive information taken into account (whether cues or clusters), versus self-insight concerning the weights given to this information. One way to address this question is to recompute the indices used in columns I and II, with exactly the same cues (properly signed, as before), but this time, without weighting them prior to summation. Columns IV and V of Table 5 present the correlations between actual and predicted judgments obtained with the indices recomputed in such a manner. The finding of interest which emerges is that the pairwise differences between these columns (see column VI) yield a picture which is practically the negative image of that found in column III. That is to say, disregarding the weights that the subjects give to the clusters leads to a greater relative loss of predictive power than in the case of individual cues. This trend suggests that the subjects, or at least some subjects, may have more self-insight about the weights they give to the concepts they use in their judgments, than to the cues which evoke them. This fact has evidently implications for the assumptions embedded in the instructions commonly given to the subjects in studies of their self-insight.

It is noteworthy that the foregoing interpretation receives some small but non-neglible support from examining the question of the tangible consequences of the information provided by the subjects about their self-insight. This issue is briefly examined below by using the subjects' self insight to see whether it helps increase the amount of explained variance obtained from the best available statistical model (see Table 1, decision rule B). One procedure to achieve this end is to add to the equations the variables reportedly having been used, but which are not in the statistically developed equations; in the case of individual cue utilization, this merely involves adding (forcing in) the cues the subjects feel they have used and which are not in the LMRs. For the concepts

inferred, the procedure requires computing interaction terms representing each cluster, forcing into the equations the cues involved in these terms which are not yet in the model, and testing whether the interaction terms add any explained variance to this recomputed base-line (Cohen and Cohen, 1975, Ch. 8).

Column VII in Table 5 shows the result of this analysis using the subjects' self-insight about individual cue utilization. The result is that in no case does the corrected measure of explained variance rise about the level previously achieved with the statistically developed model. In other words, in the occurrence the self-insight of the subjects under consideration is useless for attempting to improve on this model. In the case of clusters, on the other hand, column VIII shows that in three of the four cases at least one interaction effect does increase the corrected amount of explained variance by at least one percentage point. While this figure is admittedly low, the results are in line with the magnitudes we habitually encounter in the studies reporting success in detecting interaction effects by means of standard statistical analyses (Goldberg, 1968; Oglivie and Schmitt, 1979). In terms of the substantive process of interest, therefore, this modest finding, together with the pattern of the data exhibited by Table 5, suggests that the subjects' self insight about concept inference is likely to be grounded in reality; as noted earlier, this conclusion is by no means new (Cook and Stewart, 1975). However, the phrasing of the question by means of which it has been replicated here does have interesting implications for the problem at hand. Specifically, the view of the subjects' use of information just examined recasts in a very different light the famous discrepency between the number of cues the subjects typically report taking into account and that which is sufficient for accounting for the bulk of the explainable variance of their judgments. This becomes clear below.

B. The demand-response issue. The comparison between columns I and II of Table 6 shows that subjects attempt to infer relatively few concepts, between 3 and 5 in the present case, as opposed to an average of over 12 cues when the question they answer is put to them in terms of individual cue utlization. That is to say, the image which obtains regarding the size, and even the very existence of the discrepency noted above, is antithetically different depending upon one's choice of perspective: number of individual cues or number of concepts used. Note, moreover, that if the comparison is momentarily kept at the level of individual cues only, the subjects indicate that they use less cues when the question is phrased in terms of underlying concepts, than when it is phrased in terms of individual cues. Specifically, in contrast to an average of 12.75 cues in the latter case, they report using an average of 10 cues (if we recount a cue each time it is used in a different cluster), or an average of 8.25 cues (if those relied upon are counted only once, independently of the number of concepts on which they "load", to use an enticing analogy; see Table 6, columns I, IV and III, respectively). This latter finding is of enough importance to warrant a brief comment.

Insert Table 6 about here

There appears to be two main reasons which could explain the difference between the average number of cues found in column I (12. 75) on the one hand, and in columns III and IV (8.25 and 10, respectively), on the other. One possibility is that when asked to indicate the clusters to which cues might belong, the subjects have omitted to mention those they may have used as single measures of concepts. The other, is that the difference, especially that between 8.25 and 12.75 cues, may be indicative of the demand-response effect hypothesized earlier --even when subjects do not interpret the instructions to mean that weights should be allocated to every cue; in such a case, they could tend to interpret the instructions as a request to make an effort to allocate weights,

if not to every cue, at least to the maximum possible number of them. The data for subjects 5 and 8 are compatible with such an interpretation (see Table 6, column I versus columns III and IV).

Although the merits of the two foregoing (non-exclusive) explanations cannot be decided with the data at hand, the impressionistic evidence available from the debriefings suggests that the second explanation comes closer to describing what is actually happening.

Be this as it may, we have to contend with the fact that depending upon the instructions given, subjects report using 3 to 5 concepts in their judgments or an average of over 12 discrete cues. It could be argued that this comparison is unwarranted. Indeed, it can be held that the figure of 12.75 cues (Table 6, column I) should be compared with that of 10, or at least with that of 8.25 cues (Table 6 columns IV and III, respectively). While this is a tenable position, this remark does not affect the essence of the argument made in this section, but simply rephrases it. The reason is that this criticism implies a call for a proper comparison, a request to which it is only appropriate to respond by pointing out that to be consistent the foregoing figures should in fact be compared with those in Table 1, especially with those of equation C. In particular, the foregoing data and analyses lend support to the view that, at the level of individual cues, the correct comparison is probably between man's average of 12.75 (Table 6, column I) and that of his "C" models of 14.70 (see bottom of Table 1, first row of summary data); the similarity of the order of magnitude of these two figures hardly requires emphasis. Moreover, it is of interest to note that the figures even suggest that contrary to the prevalent imagery, the LMR procedure may be <u>less</u> discriminating with regard to the marginally relevant predictors it includes than are the human judges.

C. Transitional remarks. The crux of the foregoing discussion is that in order to be valid, the conclusion that man lacks self-insight should rest on comparisons which are internally consistent. In the light of the preceding remarks, it appears that appropriate comparisons could include the following: firstly the comparison of the factors extracted from the analysis of the matrix of cue intercorrelations of a task with the clusters obtained from answers to questions such as those illustrated above; secondly, the comparison between the "best" cues of a LMR with the subset of those selected by the subjects as also being the most important, either by their response to a direct question, or as implied by the relative weights given to the cues in the process of allocating among them the points of the scarcity scale.

To the best of my knowledge, neither of these internally consistent methods has been applied in studies of man's judgmental self-insight. In the case of objective and subjective factors, there is a practical explanation: the data required for such a comparison are rarely available. The reason is that the cue intercorrelations are typically predetermined by the researchers and factor analyzing them would produce an objective image of the clusters chosen by the experimenters, rather than one of the subjects' objective clustering policy. Rare studies where this is not the case include those by Phelps and Shanteau (1978), Holbrook(1981) and Einhorn and Koelb (1982). These researches -- none of which, incidentally, directly address the issue of self-insight -- are instructive in that they suggest that in the kind of tasks under discussion subjects may typically infer between three and six-seven concepts. By this standard, the figures obtained in column II of Table 6 are compatible with the view that the subjects may have a rather accurate perception of the number of clusters/concepts that they use in the type of judgmental tasks under consideration (although it must be noted that the difference between the findings reported by Phelps and Shanteau, 1978, and those documented in the two

other studies just referred to suggests that the number of concepts inferred may also be a function of the number of cues presented).

In the case of objectively versus subjectively identified "best" cues, however, feasibility is not an obstacle. One simply needs to set up a procedure for selecting the most important subjective cues. In the present case, the method chosen for constructing the abridged subjective indices which will be shortly analysed, takes advantage of the information available about subjects' clustering. Thus, for each respondent, the number of clusters he or she had reported to have used serves as the theoretical rationale for determining the number of individual cues to be included in the condensed index. Operationally, these cues and their subjective weights come from the responses to the probe about individual cue utilization (cf. Table 6, column I). In each case, the appropriate number of cues --inferred, as just noted, from the respondent's clustering policy, e.g. 4 for subject number four (see Table 6 column II)-- was selected by including the cues subjectively given the highest relative weights, or by random choice among the equally weighted ones in case of a tie for the last needed cue(s). The z scores of the selected cues were then properly signed, as described earlier, prior to being multiplied by their subjective weight and summed. In such a manner, the number of cues incorporated into the selectively recomputed indices range from 3 to 5, with an average of 4.25 cues per index (cf. Table 6, column II); in constrast, the standardly computed indices range from 7 to 16, with an average of 12.75 cues per index (Table 6, column I).

D. Subjects' self-insight revisited. The findings resulting from the foregoing procedure can be summarized as follows. The average squared correlation of the abridged indices just described with the actual judgments is .38. By comparison the average r^2 of the full fledged indices with the actual judgments

is .46 (cf. Table 5, column I). The observation of importance is that the ratio of these figures is .82, a result which is very close to the widely quoted conclusion of Slovic and Lichtenstein (1971, p. 684; see also Hobson, Mendel and Gibson, 1981) that a few selected cues identified by LMR analyses generally suffice to account for over 80% of the explained variance of the subjects' responses. As it turns out, these authors' conclusion sounds more unusual than it is, for the indices based on man's self-insight appear to behave in exactly the same manner. That is to say, it is possible to speculate that for many empirically developed indices 3 or 4 items might well turn out to explain most of the variance that enlarged indices might be able to explain (see supportive indication below). Be this as it may, the finding just discussed indicates that the results documented in the context of LMR analyses cannot serve as an uncritical basis for inferring that subjects lack self insight into their judgments, in particular about the number of cues they use. It is of interest to note that this conclusion is buttressed by the data of the other three subjects for whom data about individual cue utilization are also available --although without information about clustering for guiding the choice of the number of cues to include into the abridged indices. The absence of these guidelines turns out, however, to be informative. Taking the three most important cues of each subject for computing, out of necessity, uniformly abridged indices leads, indeed, to the explanation of 79% on the average of the variance of the actual judgments explained by the full fledged indices; with indices based on the four most important cues, the figures rises to 82%. Evidently, the important implication of this finding is that the indices measuring man's self-insight may be quite insensitive to the decision rule used to abridge them, (i.e. a uniform number, or one derived from other considerations, e.g. evidence about clustering), as well as to the exact number of cues selected to recompute them --in the range of roughly 3 to 5 for the type of tasks under discussion.

Recapitulation

The findings presented in this section rest on a limited data base, and must therefore be regarded as preliminary; nonetheless, they are instructive. In conjunction with the underlying argument these findings can be summarized as follows.

The subjects that we have studied assert that they utilize between 3 and 5 concepts in the fairly typical judgmental task that was used in the present study. The alternative mode of eliciting their self-insight which produced these data appears to be grounded in reality; in particular it can be used to improve the predictive power of straight LMR models and yields predictions which compare favorably with those derived from the traditional indices based on the subjects' reliance on individual cues.

The figure of 3 to 5 concepts is incompatible with the accepted view that judges tend to misperceive the extent to which they rely in their decisions on a few major variables. The likelihood that the prevalent portrayal of people's judgmental self-insight is mistaken is further buttressed by the fact that a closer examination of the way data are gathered for building the traditional indices suggests that the procedure may induce a demand-response effect. Moreover, merely by being consistent and computing predicted judgments with these standard indices exactly as one does in the case of LMRs, that is, by using selected cues according to the magnitude of their weights, shows that the "best" subjective cues appear to behave exactly as do the objective ones; in both cases 4 ± 1 cues explain about 80% or more of the variance explainable with the full fledged indices or the extended LMR's.

In short, whichever yardstick one adopts for evaluating man's self-insight -- the number of concepts inferred, or the number of cues that is sufficient for explaining the bulk of the variance explainable by the enlarged indices -- there does not appear to be serious grounds for asserting that LMR research provides evidence that man confuses all the cues he processes with the most important variables he actually takes into account in making his judgments. He may sorely lack self-insight. But, as I have endeavored to show , the results produced by the traditional type of analysis of his capability in the context of LMR judgmental tasks cannot provide the evidence necessary to establish this fact. The reason lies in part in the need to properly conceptualize what is being measured. In part it lies in the insufficient attention paid to the necessity of making comparisons which are internally consistent. And it also lies in the fact that in a world characterized by multicollinearity and monotonic relationships, most indices could well turn out to behave as do LMRs, that is, increasing either the number of index items or that of the variables entered into a LMR might well lead in both cases to rapidly diminishing returns, perhaps at comparable rates.

Be this as it may, one overall conclusion stands out --however the reader assesses each of the two measures of self-insight considered in this section and ranks them in the light of the four criteria discussed earlier. The conclusion is this: the manner in which the subjects' self-insight is commonly elicited and computed raises as many questions concerning its meaning and validity as do the values of the parameters embedded in the LMR models discussed earlier in terms of which this self-insight is appraised (cf. Part One of the analysis).

DISCUSSION

We have seen that the meaning of policy capturing by means of LMR models

turns out to be very ambiguous. This stems from the fact that there are several plausible rules for choosing between equations, none of which is compelling or standardly agreed upon by students of human judgment by means of LMRs. Often the decision rule used to select the equation(s) deemed to capture the subjects' policies is not even reported. In the judgmental task that we have considered, the degrees of freedom resulting from this situation translated into equations which could arbitrarily include from 5.5 to 14.7. cues per equation, on the average.

To compound the problem, a close examination of R^2 shows it to be a very problematic measure of the notion of cognitive control or of success in fitting a model to the judgmental data. This statement does not overstate the case for the typical range of values of n and k found in LMR research of human judgment, although it does so for a narrow range of values of R^2 near the upper limit of its maximum magnitude (Table 4). Outside this range, however, that is for the bulk of complex, real life tasks, the foregoing characterization is justified. Note, incidentally, that it is precisely under such conditions that measuring accurately cognitive consistency for providing cognitive feedback is not merely of academic interest.

Consistent with the foregoing results, it is worth noting that, stepping outside the limits of the present research, we find that across a number of published studies, R^2, n, and k are related in the predicted statistical manner. Thus, in 21 studies reviewed by Shapira (1981) for which data about the aforementioned parameters are presented, we find the following correlations: 1) between R^2 and sample size, -.52 and between R^2 and number of cues +.27; 2) between R^2 and sample size, controlling for number of cues, -.61 and between R^2 and number of cues, controlling for sample size, +.44. The anticipated artifactual relationships thus emerge as a clear trend.

Under these circumstances, it is evident that referring to the number of cues in a LMR or to the equation's R^2 as to descriptive measures of the judges' policies is far from being as enlightening as it has come to be held to be. Moreover, and although we have not touched upon these topics, three related issues make the informative values of current LMR analyses of human judgment even more problematic. The first is that cue-intercorrelations and cue redundancies affect very significantly the value and stability of beta weights (for an excellent analysis of this problem, see Gordon, 1968). The spreading representative design philosophy (Brunswik, 1955a, 1955b; Hammond and Wascoe, 1980; Hammond, MacClelland and Mumpower, 1980) of profile construction may therefore be in fundamental conflict with the methodological requirements of LMR modeling of human judgment --except when one deals with a judgmental task whose ecology is well understood and documented. The second is that whatever the quality of the estimated weights, there is no agreed upon standard way to report them. As is well known, the possibilities which include $r^2_{yi.k}$, β_i^2, $\beta_j / (\Sigma_i \beta_i)$ and $(\beta_i)(r_{yi})/R^2$ (Einhorn and Koelb, 1982; Hobson, Mendel and Gibson, 1981; Hoffman 1968), yield measures which do not necessarily rank order the cues in the same order of importance (Darlington, 1968). The third, is that if the findings pertaining to r^2 can be extrapolated to R^2 (as some empirical evidence suggests this may be the case --see Goldberg 1976, Table 1), the number of categories included in the judgmental response scale further complicates the situation by introducing another way whereby the quantity of explained variance can be arbitrarily affected to a significant degree. Thus, in the bivariate case, the effect takes the form of a systematic reduction of the explained variance; the latter shrinks increasingly as the number of response categories diminishes and as the size of r^2 grows. To illustrate, the same relationship which would yield a value of $r^2 = .65$ with a five point response-scale, will produce one of $r^2 = .76$ with a ten point response-scale, or vice versa (see

Martin, 1973, Table 1). This effect which compares in magnitude with those previously documented but which is independent of them, is therefore potentially strictly additive. Moreover, its impact reaches its maximum as the amount of explained variance approaches its limit, that is in the range of values where on the basis of Table 4 one might have concluded that because of the reduced effect of k and n on the explained variance, it becomes relatively safe to relate to R^2 as to a measure of cognitive control or of success in capturing a judge's policy. Obviously, this restricted assumption, too, is unsafe.

Another result further obscures the meaning of current LMR models of policies. This is the finding that the assumption that subjects can be characterized by one overtime policy may be questionable. We have seen that the equations developed on the whole object sample and those developed on sequential subsets of it exhibit differences which include the variables that characterize the policies, their weights, signs, and the amount of (adjusted) explained variance already noted.

That is to say, not only are there problems of measurement, but the very notion of policy that LMRs presumably capture turns out to be elusive. What is the proper amount of profiles and/or time spent judging them which yields a meaningful image of a subject's policy, or at least reflects a natural segment of it? Evidently, what the analysis of data gathered during a typical experimental session produces is often a statistical average which needs not bear a direct resemblance to any of the policies involved, in particular the latest one being implemented. Under such circumstances, the yardstick used for supplying cognitive feedback to the subjects is obviously very problematic.

For the purpose of assessing the judges' self-insight, the foregoing difficulties which are inherent in the nature of LMR models of human judgment, are compounded by those created by the manner in which data about self-insight are elicited. In particular, the evidence is that the findings documented to

date involve a demand-response effect. Thus, taking the subjects' claim that they relate to clusters of cues rather than to individual ones as a working hypothesis, one finds that in the present judgmental task their intuition is that they utilized between 3 and and 5 conceptual variables in their decisions. These values are at variance with the image of people being unaware of the extent to which they rely in their judgments on a few major variables. Moreover, constructing traditional indices which are simply consistent with the LMR's computations with which they are to be compared, reproduces the famous LMR finding, namely, that a few selected cues suffice to account for about 80% or more of the variance explained by the full-fledged set of cues.

In sum, the grounds and rationale for building LMR policy-capturing models and for providing subjects with feedback to improve their cognitive awareness and/or cognitive control turn out to be questionable in the extreme. This holds true in terms of the objective model, in terms of the subjective data gathered about self-insight, and with regard to the manner in which the two are then compared. In light of this situation, it is inescapable that the original aim of policy-capturing cannot be said to have been achieved. This aim was stated by one of its pioneers to be the confrontation of the problem from which all others may be held to stem, namely, that judgment is a process that we cannot trace:

> "It is as if we put our empirical data into a computing machine, the processes of which we did not understand and which frequently produced different results depending on which machine we used and when we used it." (Hammond, 1955, p. 255).

This characterization of judges applies evidently as well, if not better, to the proposed solution --the present day LMR models. Indeed, there is little doubt that for many applications, we have replaced one black box by another. It is intriguing, therefore, to observe that the use of LMR models for cognitive feedback is spreading (Hammond, Rohrbaugh, Mumpower and Adelman, 1977), and

appears to be well received by the subjects, sometimes with impressive results (Hammond and Adelman, 1976, Anderson et. al. 1981). On the background of the current shortcomings of the approach that we have discussed, weaknesses which take their full meaning in the light of the explicit theoretical disavowals noted in the Introduction, one is puzzled by the situation which has developed. It could be that its explanation lies in the subjects' favorable reaction that it is tempting to regard as an implicit indication that something is fundamentally right in the endeavor, in spite of all the present arguments to the contrary. But it is clear that this reaction of the subjects should be checked for a possible artifact which appears to have been completely overlooked in the literature --despite its being a familiar one. This artifact can be operationalized by the following questions: How would the subjects react to, and accept, randomly generated models substituted for theirs? Or, similarly, how would subjects relate to their own models, if, adapting one of Milgram's (1974) research designs, they were presented to them not at the terminal of a computer, and in an academic context, but in a less authoritative environment, and without the computer aura? Without evidence to the contrary, it is hard to escape the feeling that what we may presently be witnessing is the effect of the principle that for many subjects a sufficiently sophisticated methodology is indistinguishable from magic, to paraphrase an aphorism quoted by Parker (1976, p. 1).

There is a whole literature dealing with models developed for the purpose of predicting an objective criterion, rather than for that of reproducing or capturing a judge's policy on which our discussion has focused. Many of the issues that we have raised have been addressed in this literature. In particular, the question of the sample size and of the number of predictors in an equation have been discussed or noted by Einhorn and Hogarth (1975; 1982), Morans and Drasgow (1978), and Keren and Newman (1978). Similarly, the problem of the instability of the beta weights has been stressed in this context by many

researchers since Darlington (1968), including the authors just referred to, Schmidt (1972), Schmitt and Levine (1977), and Schoemaker and Waid (1982). With regard to R^2 Cattin (1980) has recommended the general use of a corrected measure of this coefficient and suggested a more accurate way of computing R^2 when n 50. The issue of self-insight, too, has been investigated in terms of the subjects' ability to predict an objective criterion. To date, the evidence for this kind of self-insight is mixed, with findings which sometimes reflect favorably on man's ability, and sometimes less so (Schmitt and Levine, 1977; Schmitt, 1973; Gray, 1979; Shoemaker and Waid, 1982), in part, perhaps, because of the effect of the values of n and k used in the models (Cattin, 1980, p. 413).

The implications of the findings and warnings found in this literature have not been generalized to the activity of policy-capturing and insight determination, however. Modeling policies is therefore an endeavor which currently not only lacks a theoretical justification, but which also involves many serious practical problems. The result is that policy capturing is presently an activity with a very questionable rationale. The quality of cognitive feedback given to the subjects is anyone's guess. And whether or not man has self-insight into the policies he applies in his judgments remains a cluttered and unsettled issue, despite the pivotal importance of this question for some research (Stillwell, Seaver and Edwards, 1981).

From a remedial perspective, some of the problems that we have considered are procedural, e.g. number of response categories in the judgmental scales, replacement of R^2 by a measure corrected for degree of freedom, manner in which data about self-insight are elicited and analyzed. Others are inherent in the LNR methodology, for instance the fact that --to use Kerlinger and Pedhazur's (1973, p. 442) words-- "A serious weakness of multiple regressions is what can be called the unreliability of regression weights."

The first class of problems can be readily dealt with without great difficulty, for independently of whether or not they have a solution in some deep sense, they can be fairly effectively controlled by holding their effect constant. On the other hand, the second category of problems requires serious analytical and methodological work which may take years to bear fruit. It is possible for some time to "look at the other side of the coin", to continue quoting Kerlinger and Pedhazur (1973, p. 444 ff.), for LMR's have indeed strengths which could be taken advantage of, if proper caution is exercised. In the long run, however, it seems clear that if LMR modeling of human judgment for policy-capturing purposes is to be regarded as a justified endeavor, the validity of the models as a reflection of the judges' policy must be demonstrable. Here the likely equivalence of various types of weights for predictive purposes clearly defines the problem that must be confronted: in the light of the current inability of policy-capturing models to effectively compete with these alternative schemes in terms of relative (and at times absolute) levels of achievement, one of the two following alternatives must evidently be faced. The first is to convincingly capture man's judgmental policies, and thus show that what is offered as an aid to the limitations of his self-insight is valid feedback, and not some mixture of elements of a policy with statistical and methodological artifacts, the latter unextricably intertwined with, and possibly completely overshadowing the former. The other is to accept the prospect that the endeavor may increasingly be seen as devoid of a defensible scientific justification.

The seeds of this conclusion are implicitly found in the points made in the literature on predictive linear models to which we have referred above. These points have tended to be discussed as discrete issues, however, a fact which explains perhaps the lack of sufficient attention paid to them for the

topic at hand. This assumption underlies the present attempt to bring them together and to spell out their implications for current policy-capturing work. Looking back at the picture which emerges from the discussion, it is not far-fetched to say that these implications could be regarded with some justification by a critic of policy-capturing work as suggesting that the king is currently naked.

Lest this remark be misunderstood, let me hasten to add that I emphatically do not believe that this is the case. I am convinced that policy-capturing research is theoretically important, and that the aim of providing subjects with cognitive feedback is of practical significance. More importantly, past studies which have avoided some or all of the pitfalls that we have discussed have made significant contributions to our knowledge. The point of the foregoing remark is to call attention to the fact that once this has been said there is, however, only so much which can be accomplished without confronting the key problems involved in the LMR modeling approach. It is in this perspective that it is important to realize that the list of threats to the validity of current policy-capturing work and its applications create a situation which is not a strategic one for complacently continuing carrying out research in the habitual way, and thus risking the prospect of having to face the accusation just noted. Put differently, the insights and findings accumulated to date clearly need to be scrutinized and consolidated in the light of the potential threats to their validity that we have discussed.

SUMMARY AND CONCLUSION

We have considered a number of problems which contitute validity threats to current LMR models of human judgments. These include:

1. The lack of unambiguous criteria for including variables in a LMR model and for choosing the equation deemed to reflect the judges' policy, a problem which creates an unstructured situation that needs to be given attention and corrected.

2. The necessity of agreeing on an operationalization of the notion of weight, and of presenting evidence of their stability; at least that of reporting the weights in a manner which permits one to recompute them differently within the LMR paradigm.

3. The need of documenting the effect on R^2 of the number of categories in the response scales given to the subjects.

4. The necessity of clarifying what is meant by cognitive consistency and by success in modeling a judge's policy. In particular, the need of examining whether $\hat{R}^2$ should replace R^2 as a standard measure of these notions. And, if so, which method should be used to correct for the biases involved in the stepwise development of equations

5. The necessity of determining the appropriate sample size for model(s) building, in light of the evidence that subjects may systematically change their policies in the process of judging a set of profiles.

And, after clarifying and improving the models along such lines, and in the process increasing their claim to trustworthiness,

6. The need of eliciting the subjects' self-insight according to a theoretical conceptualization of the process investigated. At least, that of eliciting the relevant data in a manner which does not induce a demand-response effect, and of computing the predictions derived from self-insight in a fashion which parallels in its logic the procedure used with LMR models.

In sum, one of the central problems of LMR models of human judgment is that for the typical values of n and k used for model building, regression weights are unreliable. Their stability, however, is of crucial importance for the tenability of the assumption that a policy has been captured and deserves, therefore, to be fed-back to the subjects. With the beta weights defaulting, a remaining indicator of the quality and stability of a model is potentially the value of R^2, as we find, indeed, that this coefficient is used in the literature --in general implicitly, but on occasion explicitly (Hammond and Marvin, 1981). However, we have seen that this coefficient can be seriously misleading for a number of reasons. These include the object sample size used for model development, the number of variables included in the model, the number of categories in the response scale, etc. The interpretation of current LMR models of human judgment is therefore in many cases uncompelling. This situation is compounded by problems in the manner in which the subjects' self-insight is elicited and analyzed. Together these difficulties raise fundamental questions about the accuracy of the characterization of human information processing derived from this evidence, about the validity of the portrayal of the subjects' self-insight it sustains, and about the usefulness of the feed-back provided to the subject.

This situation needs evidently to be remedied. By spelling out the extent to which it is problematic, this paper will hopefully contribute to the stimulation of the necessary corrective work. In the meantime, it should serve as a warning against accepting with too much faith some of the conclusions which stem from current LMR models uncritically related to and used as dependable descriptions of human judgmental policies.

FOOTNOTES

1. This holds true for most decision rules owing to the fact that we can encounter situations where R^2 is statistically significant, while none of the tests for the individual X's are, and conversely, situations where the t tests for one or more individual predictors are statistically significant, while the overall R^2 is not (see Cohen and Cohen, 1975, section 3.7, especially pp. 108-109). Not surprisingly, proposals to safely deal with these problems are open to the criticism that they are overly conservative.

2. One problem in comparing the R^2's of independent equations is that the significance of differences between amounts of explained variance is not readily determinable. One common heuristic procedure under the circumstances is to regard a difference of 1% of explained variance as noteworthy. However, many researchers often seem to interpret this rule of thumb as meaning that such an amount is noticeable, rather than necessarily of substantive importance. As a result, there is a zone of ambiguity in the interpretation of the significance of a gain/loss of explained variance which extends at times to 2-3%. Workers specializing in the use of LMRs often resolve it (especially for large values of R^2) by applying the following principle: a difference equal to or larger than 10% of the quantity $1-R^2$ is regarded as "significant". In the case at hand, the values of the R^2's under consideration spread around .70; consequently, $(1-R^2)/10 \approx 3\%$. By this criterion, a difference of approximately this size, or greater, between two R^2's may be regarded as being unambiguously "significant". These guidelines are offered with no stronger claim for them than the fact that they are commonly used and may be useful to fix ideas.

3. In practical terms this task can be carried out in two ways. The first approach is to get continuous data on the process, a task which turns out to be very difficult and cumbersome, owing to both problems of data recording and analysis. Note that this difficulty is also encountered when data are gathered about individual cues; this has led in this case to the current use of the scarcity scale method referred to above. In particular, the subjects are requested to allocate points in retrospect, i.e. the data elicited are about the weights of the cues psychologically averaged after the fact across profiles. This method is used despite its imperfections, both substantive and procedural (cf. Ericsson and Simon, 1980), the overriding consideration being its practicality, and the supporting rationale that in matters of policy, there is also a substantive interest in the validity of insights as recollected and communicated in retrospect. Along a similar line of reasoning, it can be argued that if the information about clusters is conceptually important and different from that about individual cues, and if this difference is observable and robust, the application in this case as well of the second (retrospective) approach just noted could yield informative results--despite its acknowledged imperfections. At the very least, the results would be directly comparable with those obtained about individual cues. In a nutshell, this is the rationale on which the forthcoming analyses rest.

4. Because, as we have seen, cue signs may change according to the data base used for model building, this procedure assumes that the sample size used in Table 1 is an appropriate one for the purpose at hand, an assumption which is of course open to question. However, the implication of interest of the findings to be discussed turns out to be independent of this assumption. This will become clear as the more general analysis and argument presented later will show.

REFERENCES

Anderson B.F. et al.

1981 Second Report to the Rocky Flats Monitoring Committee Concerning Scientists' Judgments of Cancer Risk. Center for Research on Judgment and Policy, Report No. 233, University of Colorado.

Brehmer B.

1978 "Response Consistency in Probabilistic Inference Tasks." Organizational Behavior and Human Performance. Vol. 22, pp. 103-115.

Brehmer B., R. Hagafors and R. Johansson

1980 "Cognitive Skills in Judgment: Subjects' Ability to Use Information About Weights, Function Forms and Organizing Principles." Organizational Behavior and Human Performance. Vol. 26, pp. 373-385.

Brehmer B. and J. Kuylenstierna

1980 "Content and Consistency in Probabilistic Inference Tasks." Organizational Behavior and Human Performance, Vol. 26, pp. 54-64.

Brehmer B. and K.R. Hammond

1977 "Cognitive Factors in Interpersonal Conflict". In D. Druckman (ed.) Negotiations: Social Psychological Perspectives. Beverly Hills, California: Sage Publications.

Brehmer B. and G. Quarnstrom

1976 "Information-Integration and Subjective Weights in Multiple-Cue Judgments." Organizational Behavior and Human Performance, Vol. 17, pp. 118-126.

Brunswik E.

1955(a) "Representative Design and Probabilistic Theory in a Functional Psychology." Psychological Review, Vol. 62, pp. 193-217.

Brunswik E.

1955(b) "In Defense of Probabilistic Functionalism: A Reply." Psychological Review, Vol. 62, pp. 236-242.

Bucuvalas M.J.

1978 "The General Model and the Particular Decision: Decision Makers' Awareness of their Cue Weightings." Organizational Behavior and Human Performance, Vol. 22, pp. 325-349.

Camerer C.

1981 "General Conditions for the Success of Bootstrapping Models." Organizational Behavior and Human Performance, Vol. 27, pp. 411-422.

Cattin P.

1980 "Estimation of the Predictive Power of a Regression Model." Journal of Applied Psychology, Vol. 65, pp. 407-414.

Cohen J and P. Cohen

1975 Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Cook R.L. and Stewart T.R.

1975 "A Comparison of Seven Methods for Obtaining Subjective Descriptions of Judgmental Policy." Organizational Behavior and Human Performance, Vol. 13, pp. 31-45.

Darlington R.B.

1968 "Multiple Regression in Psychological Research and Practice." Psychological Bulletin, Vol. 69, pp. 161-182.

Dawes R.M.

1979 "The Robust Beauty of Improper Linear Models in Decision Making". American Psychologist, Vol. 34, pp. 571-582.

Dawes R.M.

1971 "A Case Study of Graduate Admissions: Application of Three Principles of Human Decision Making". American Psychologist, Vol. 26, pp. 180-188.

Dawes R.M. and B. Corrigan

1974 "Linear Models in Decision-Making". Psychological Bulletin. Vol. 81, pp. 95-106.

Dorans N. and F. Drasgow.

1978 "Alternative Weighting Schemes for Linear Prediction." Organizational Behavior and Human Performance, Vol. 21, pp. 316-345.

Dudycha L.W. and J.C. Naylor

1966 "Characteristics of the Human Inference Process in Complex Choice Behavior Situations." Organizational Behavior and Human Performance, Vol. 1, pp. 110-128.

Einhorn H.J.

1971 "The Use of Non-Linear, Non-Compensatory Models As A Function of Task And Amount of Information." Organizational Behavior and Human Performance, Vol. 6, pp. 1-27.

Einhorn H.J.

1970 "The Use of Non-Linear, Non Compensatory Models in Decision-Making." Psychological Bulletin, Vol. 73, pp. 221-230.

Einhorn H.J. and R.M. Hogarth

1982 "Prediction, Diagnosis, and Causal Thinking in Forecasting." Journal of Forecasting, Vol. 1, pp. 1-14.

Einhorn H.J. and C.T. Koelb

1982 "A Psychometric Study of Literary-Critical Judgment." Modern Language Studies, in press

Einhorn H.J., D.N. Kleinmuntz and B. Kleinmuntz

1979 "Linear Regression and Process-Tracing Models of Judgment". Psychological Review, Vol. 86, pp. 465-485.

Einhorn H.J. and R.M. Hogarth

1975 "Unit Weighting Schemes for Decision Making." Organizational Behavior and Human Performance, Vol. 13, pp. 171-192.

Ericsson K.A. and H.A. Simon

1980 "Verbal Reports as Data." Psychological Review, Vol. 87, pp. 215-251.

Goldberg L.R.

1976 "Man Versus Model of Man: Just How Conflicting Is That Evidence?" Organizational Behavior and Human Behavior, Vol. 16, pp. 13-22.

Goldberg L.R.

1970 "Man Versus Model of Man: A Rationale, Plus Some Evidence, For A Method of Improving On Clinical Inferences". Psychological Bulletin, Vol. 73, No. 6, pp. 422-432.

Goldberg L.R.

1968 "Simple Models Or Simple Processes? Some Research On Clinical Judgment." American Psychologist, Vol. 23, pp. 483-496.

Gordon R.A.

1968 "Issues in Multiple Regression." American Journal of Sociology, Vol. 73, pp. 592-616.

Gray C.W.

1979 "Ingredients of Intuitive Regression." Organizational Behavior and Human Performance, Vol. 23, pp. 30-48.

Green P.E. and D.S. Tull

1970 Research for Marketing Decisions. Second Edition. Englewood Cliffs, N.J.: Prentice Hall.

Hammond K.R.

1955 "Probabilistic Functioning and the Clinical Method." Psychological Review. Vol. 62, pp. 255-262.

Hammond K.R. and B.A. Marvin

1981 Report to the Rocky Flats Monitoring Committee Concerning Scientists' Judgments of Cancer Risk. Center for Research on Judgment and Policy, Report No. 232, University of Colorado.

Hammond K.R. and N.E. Wascoe (eds.)

1980 New Directions for Methodology of Social and Behavioral Science: Realizations of Brunswik's Representative Design. San Francisco: Jossey-Bass.

Hammond K.R., G.H. McClelland and J. Mumpower

1980 Human Judgment and Decision Making: Theories, Methods and Procedures. New York, N.Y.: Praeger.

Hammond K.R., J. Rohrbaugh, J. Mumpower and L. Adelman

1977 "Social Judgment Theory: Applications in Policy Formation". In M.F. Kaplan and S. Schwartz (eds.) Human Judgment and Decision Processes in Applied Settings. N.Y.: Academic Press.

Hammond K.R. and L. Adelman

1976 "Science, Values and Human Judgment." Science, Vol. 194, pp. 389-396.

Hammond K.R., T.R. Stewart, B. Brehmer and D.D. Steinmann

1975 "Social Judgment Theory". In M.F. Kaplan and S. Schwartz (eds.) Human Judgment and Decision Processes. New York, N.Y.: Academic Press.

Hammond K.R. and B. Brehmer

1973 "Quasi-Rationality and Distrust: Implications for International Conflict". In L. Rappoport and D.A. Summers (eds.) Human Judgment and Social Interaction. N.Y.: Holt, Rinehart and Winston.

Hobson C.J., R.M. Mendel and F.W. Gibson

1981 "Clarifying Performance Appraisal Criteria." Organizational Behavior and Human Performance. Vol. 28, pp. 164-168.

Hoffman P.J.

1968 "Cue Consistency and Configurality in Human Judgment." In. B. Kleinmuntz (ed.) Formal Representations of Human Judgment, pp. 53-90. New York. N.Y.: John Wiley and Sons.

Hoffman P.J.

1960 "The Paramorphic Representation of Clinical Judgment". Psychological Bulletin, Vol. 57, pp. 116-131.

Hogarth R.M.

1980 Judgment and Choice. N.Y.: John Wiley and Sons.

Holbrook M.B.

1981 "Integrating Compositional and Decompositional Analyses to Represent the Intervening Role of Perceptions in Evaluative Judgments." Journal of Marketing Research, Vol. XVIII, pp. 13-23.

Keren G. and J.R. Newman

1978 "Additional Considerations with Regard to Multiple Regression and Equal Weighting." Organizational Behavior and Human Behavior, Vol. 22, pp. 143-164.

Kerlinger F.N. and E.J. Pedhazur

1973 Multiple Regression in Behavioral Research. New York, N.Y.: Holt, Rinehart and Winston.

Lane D.M., K.R. Murphy and T.E. Marques

1982 "Measuring the Importance of Cues in Policy Capturing," Organizational Behavior and Human Performance, Vol. 30, pp. 231-240

Libby R.

1976(a) "Man Versus Model of Man: Some Conflicting Evidence." Organizational Behavior and Human Performance. Vol.16, pp. 1-12.

Libby R.

1976(b) "Man Versus Model of Man: The Need for a Nonlinear Model." Organizational Behavior and Human Performance, Vol. 16, pp. 23-26.

Martin W.S.

1973 "The Effects of Scaling on the Correlation Coefficient: A Test of Validity." Journal of Marketing Research. Vol. X, pp. 316-318.

Meehl P.E.

1954 Clinical Versus Statistical Prediction. Minneapolis, Minnesota: University of Minnesota Press.

Milgram, S.

1974 Obedience to Authority: An Experimental View. New York, N.Y.: Harper and Row.

Nie N.H. et.al.

1975 SPSS. Second Edition. New York, N.Y.: McGraw-Hill Book Co.

Ogilvie J.R. and N. Schmitt

1979 "Situational Influences on Linear and Nonlinear Use of Information." Organizational Behavior and Human Performance, Vol. 23, pp. 292-306.

Parker D.B.

1976 Crime by Computer. New York, N.Y.: Charles Scribner's Sons.

Phels R.H. and J. Shanteau

1978 "Livestock Judges: How Much Information Can An Expert Use?" Organizational Behavior and Human Performance, Vol. 21, pp. 209-219.

Schmidt F.L.

1972 "The Reliability of Differences Between Linear Regression Weights in Applied Differential Psychology." Educational and Psychological Measurement, Vol. 32, pp. 879-886.

Schmitt N.

1978 "Comparison of Subjective and Objective Weighting Strategies in Changing Task Situations." Organizational Behavior and Human Performance, Vol. 21, pp. 171-188.

Schmitt N. and R.L. Levine

1977 "Statistical and Subjective Weights: Some Problems and Proposals." Organizational Behavior and Human Performance, Vol. 20, pp. 15-30.

Schoemaker P.J.H. and C.C. Waid

1982 "An Experimental Comparison of Different Approaches to Determining Weights in Additive Utility Models." Management Science, in press

Shapira M.

1981 A Study of Information Utilization in Disposition Assignments of Probation Officers. Unpublished Ph.D. dissertation, The Hebrew University of Jerusalem.

Slovic P., B. Fischhoff and S. Lichtenstein

1977 "Behavioral Decision Theory". Annual Review of Psychology, Vol. 28, pp. 1-39.

Slovic P. and S.C. Lichtenstein

1971 "Comparison of Bayesian and Regression Approaches to the Study of Information-Processing in Judgment". Organizational Behavior and Human Performance. vol. 6, pp. 629-744.

Stillwell W.G., D.A. Seaver and W. Edwards

1981 "A Comparison of Weight Approximation Techniques in Multiattribute Utility Decision Making." Organizational Behavior and Human Performance. Vol. 28, pp. 62-77.

Tucker L.R.A.

1964 "A Suggested Alternative Formulation in the Developments by Hursch, Hammond and Hursch, and by Hammond, Hursch and Todd." Psychological Review, Vol. 71, pp. 528-530.

Wonnacott R.J. and T.H. Wonnacott

1979 Econometrics. Second edition. New York, N.Y.: John Wiley and Sons.

Table 1

Variations in Judgmental Policies (Cue utilization and beta weights) According to three Decision-Rules for Model Building

Subject number	Decision Rule**	Cues* a.	b.	c.	d.	e.	f.	g.	h.	i.	j.	k.	l.	m.	n.	o.	p.	r^2_m	Adjusted R^2***	Number of Cues in Equation
1	A	-.19			.24	.60		-.36		.29	.20	.27						.67	.64	7
	B	-.18			.26	.64		-.36		.24	.28	.30			.14		-.13	.70	.66	9
	C	-.14	-.15	.17	.29	.64	.06	-.28	-.07	.16	.34	.33	-.11	.18	.23	-.07	-.11	.74	.67	16
2	A				.20	.56	.23		.32									.69	.67	4
	B				.23	.56	.22	-.21	.17									.71	.69	5
	C	.07	.05		.23	.55	.21	-.24	.11	.18	-.07	.03	.16	.04	-.06	-.08	-.03	.75	.68	15
3	A				.69	.29				.21		.28	.22	.30				.78	.76	6
	B				.69	.29				.21		.28	.22	.30				.78	.76	6
	C	.03	.06	.03	.68	.27		-.10	-.03	.26		.27	.20	.29	-.05	.11		.80	.75	11
4	A		-.15		.30	.41			.15			-.27	.15				-.16	.74	.71	7
	B		-.15		.30	.41			.15			-.27	.15				-.16	.74	.71	7
	C	.03	-.11	-.10	.27	.41	.05	-.03	.28	.02	.07	-.35	.18	-.19	-.09	.01	-.22	.76	.70	14
5	A		-.27		.24	.40					.19			.34				.32	.49	5
	B	-.15	-.18		.24	.42		-.14			.20			.28	.14			.57	.52	8
	C	-.11	-.22	.10	.23	.43	.09	-.14		-.06	.29	-.03		.26	.18	-.02	-.09	.59	.49	14
6	A				.40				.24									.24	.22	2
	B				.28			.30	.44			-.12	.14					.30	.25	5
	C	-.03	.14	-.04	.31			.20	.38	.13	-.10	-.08	.17	.02		.08	-.06	.33	.18	13
7	A		-.21			.62			.20				.29					.63	.61	4
	B		-.20		-.17	.72			.23		.12		.32					.66	.63	6
	C	-.02	-.18	.08	-.16	.76	-.02	.06	.19		.14	.08	.31	.11	.17	-.04	.03	.69	.60	15
8	A		-.31		.35	.33	.14	.21	.24	-.22	.17	-.36						.79	.76	9
	B		-.28		.31	.33	.18	.21	.36	-.22	.19	-.42		-.17				.80	.77	10
	C	-.05	-.27	-.04	.30	.34	.18	.19	.39	-.19	.18	-.42	.09	-.22	-.06	-.08	-.05	.82	.77	16

Summary Data	Decision Rule: A	B	C
Average number of cues in equations	5.5	7.0	14.7
Average R^2	.632	.657	.685
Average Adjusted R^2	.607	.624	.605

* (a) sex; (b) age; (c) ethnic origin; (d) I.Q.; (e) high school graduation grade; (f) socio-economic background; (g) marital status; (h) health; (i) achievement expectations; (j) nature of relations with high school teachers; (k) time spent doing homework during last year of high school; (l) fear of failure; (m) living expenses arrangements; (n) political activities; (o) social connections with university staff members; (p) sociability.

** Decision rules for model building: by inclusion in the equation of (A) the variables with a significant beta weight at the .05 level or beta only; (B) all the variables which contribute at least 1% of explained violence to the equation; (C) all the variables which contribute any measurable amount of explained variance to the equation.

*** See text for details of adjustment.

Table 2

Subjects' Judgmental Policies as Reflected in the object samples

split into two Sequential Subsamples

Subject Number	Number of Samples**	Cues* a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	R^2	Adjusted R^{2}***	Number of lines in Equation
1	I	-.22	-.29	.16		.64		-.48			.44		-.16		.12		-.26	.82	.76	9
	II					.92			.23				.10					.90	.89	3
	B	-.18			.26	.64		-.36		.24	.28	.30			.14		-.13	.70	.66	9
2	I				.32	.51		-.18					.13	.16		-.11		.73	.67	6
	II					.65	.34		.46									.81	.79	3
	B				.23	.56	.22	-.21	.17									.71	.69	5
3	I	.12		-.10	.32	.41	.13	-.15	.15	.17			.33		-.25	.14		.84	.76	11
	II			.16	.66	.28								.31				.82	.79	4
	B				.69	.29				.21		.28	.22	.30				.78	.76	6
4	I		-.28		.32	.38			.17	-.14	.19	-.34					-.19	.83	.78	8
	II		.18	-.14	.60	.41		-.22		.48		.15	.21			.12	-.18	.81	.73	10
	B		-.15		.30	.41			.15			-.27	.15				-.16	.74	.71	7
5	I	-.11	-.22		.30	.42	.15		.23		.29	-.26	-.13				-.23	.79	.70	10
	II	-.13				.46		-.32			.21		.20	.32	.16			.64	.55	7
	B	-.15	-.18		.24	.42		-.14			.20			.28	.14			.57	.52	8
6	I		.30		.64		-.12	.32	.28	.38		.38		.41	.21	.17		.49	.29	10
	II	-.14	-.14	.18	.58		.22						.21	.20	.29	-.16	-.19	.71	.60	10
	B				.28			.30	.44			-.12	.14					.30	.25	5
7	I		-.14		.23	.57			-.11	.20		.46	.39	.31	.34		.24	.72	.61	10
	II	-.15	-.13		-.19	.83			.33		.19		.24				-.12	.76	.69	8
	B		-.20		-.17	.72			.23		.12		.32					.66	.63	6
8	I		-.45		.38	.43		.23	.19	-.33	.24	-.22						.91	.88	8
	II	-.19			.54	.21	.29		.18	.22	.12	-.19						.84	.79	8
	B		-.28		.31	.33	.18	.21	.36	-.22	.19	-.42		-.17				.80	.77	10

Summary Data

	Nature of Sample: I	II	I+II	B
Average R^2	.77	.79	.78	.66
Average Adjusted R^2	.68	.73	.70	.62
Average number of cues in equations	9	6.6	7.8	7

* See key in Table 1

** I, Sequential profiles 1-36; II, Sequential profiles 37-72; B, whole sample of 72 (Taken from Table 1)

*** See text for details of adjustment

TABLE 3

Summary Data of Subjects' Judgmental Policies As Reflected in 1) The Randomly Split Object Samples 2) The Sequentially Split Object Samples 3) The whole Object Sample (All the equations in accordance with modeling decision-rule B in Table 1)

	(I) Average in Randomly Split Subsamples			(II) Average in Sequentially Split Subsamples (see Table 2)			(III) Value in whole Sample (see Table 1)		
Subject Number	R^2	Adjusted R^2	Number of variables in equation	R^2	Adjusted R^2	Number of variables in equation	R^2	Adjusted R^2	Number of variables in equation
1	.78	.72	7.5	.86	.83	6.0	.70	.66	9.0
2	.77	.71	6.5	.77	.73	4.5	.71	.69	5.0
3	.82	.77	6.5	.83	.78	7.5	.78	.76	6.0
4	.78	.74	6.5	.82	.75	9.0	.74	.71	7.0
5	.68	.60	6.5	.71	.62	8.5	.57	.52	8.0
6	.51	.33	8.5	.60	.44	10.0	.30	.25	5.0
7	.71	.65	6.0	.74	.65	9.0	.66	.63	6.0
8	.80	.77	5.0	.88	.84	8.0	.80	.77	10.0
Average Across Subjects	.73	.66	5.5	.78	.70	7.8	.66	.62	7.0

TABLE 4

Values of $\overset{*2}{R}$'s for Selected Values of R^2, and combinations of k and n.

	$\overset{*2}{R}$ = .30				$\overset{*2}{R}$ = .40				$\overset{*2}{R}$ = .50			
n \ k	5	6	7	8	5	6	7	8	5	6	7	8
20	.48	.52	.56	.59	.56	.59	.62	.65	.63	.66	.68	.71
30	.42	.44	.47	.49	.50	.52	.54	.57	.59	.60	.62	.64
40	.39	.41	.43	.44	.48	.49	.51	.52	.56	.58	.59	.60
50	.37	.39	.40	.41	.46	.47	.49	.50	.55	.56	.57	.58
	$\overset{*2}{R}$ = .60				$\overset{*2}{R}$ = .70				$\overset{*2}{R}$ = .80			
n \ k	5	6	7	8	5	6	7	8	5	6	7	8
20	.71	.73	.75	.77	.78	.79	.81	.83	.85	.86	.87	.88
30	.67	.68	.70	.71	.75	.76	.77	.78	.83	.84	.85	.86
40	.65	.66	.67	.68	.74	.75	.75	.76	.83	.83	.84	.84
50	.64	.64	.66	.67	.73	.74	.74	.75	.82	.82	.83	.83

Table 5

Relationship Between Actual Judgments And Self-Insights About the Cues And the Clusters Used in Making These Judgments

1 = Actual Judgments; 2 = Judgments Predicted From Self-Insight About The Individual Cues Used And Their Weights; 3 = Judgments Predicted From Self-Insight About Inferred Clusters And Their Weights; 4 = Judgments Predicted From Self-Insight About Individual Cues Used (unweighted); 5 = Judgments Predicted From Self-Insight About Inferred Clusters (unweighted).

Subject Number	(I) $R^2_{1,2}$ (Squared Correlation between actual judgments & judgments predicted from self-insight about individual cues & their weights)	(II) $R^2_{1,3}$ (Squared Correlation between actual judgments & judgments predicted from self-insight about inferred clusters & their weights)	(III) Direction of difference between columns II-I	(IV) $R^2_{1,4}$ (Squared Correlation between actual judgments & judgments predicted from self-insight about unweighted individual cues)	(V) $R^2_{1,5}$ (Squared Correlation between actual judgments & judgments predicted from self-insight about unweighted clusters)	(VI) Direction of difference between columns V-IV	(VII) Increase in R^{*2} obtained from self-insight about individual cue utilization *	(VIII) Increase in R^{*2} obtained from self-insight about clusters inference *
4	.45	.55	+	.51	.38	-	.00	.02
5	.36	.40	+	.40	.39	-	.00	.01
7	.38	.43	+	.37	.34	-	.00	.00
8	.63	.56	-	.61	.61	=	.00	.01

* Decreases in R^{*2} are reported as zero gain.

TABLE 6

Number of cues and concepts used in Judgments

(Data from Self-Insight)

Subject Number	Number of Cues Reported Having Been Used	Number of Clusters Identified	Number of Cues In Clusters:	
			Counted Once	Counted as Many times As Used
	I	II	III	IV
4	16	4	7	9
5	12	5	7	11
7	16	3	14	15
8	7	5	5	5
Average	12.75	4.25	8.25	10